ICMR 2017 Mini-Tutorial

TRECVid Ad hoc Video Search

Georges Quénot
Multimedia Information Modeling and Retrieval Group

Laboratoire d'**I**nformatique de **G**renoble

with input from George Awad (Dakota Consulting, Inc and NIST)
and many others

June 6, 2017

Tutorial Outline

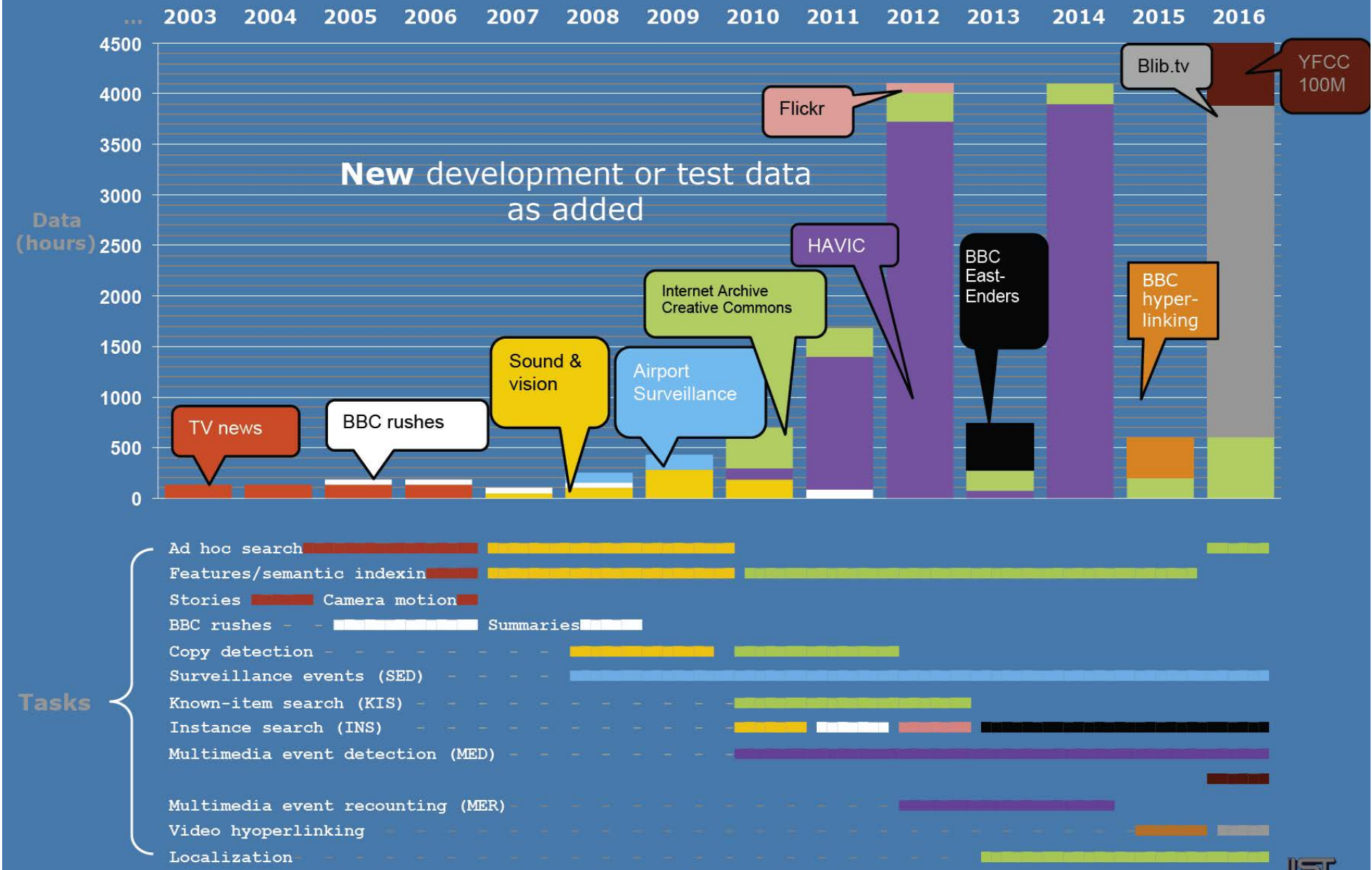
- Part I: the Ad hoc Video Search (AVS) task
- PART II: some participants' implementations

Part I
the Ad hoc Video Search (AVS) task

G. Awad, J. Fiscus, D. Joy, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, M. Eskevich, R. Aly, R. Ordelman, G. J. F. Jones, B. Huet, M. Larson. ***TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking.*** TRECVID 2016, NIST, USA.

http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/tv16overview.pdf

TRECVID's Evolution



Ad-hoc Video Search Task Definition

- Goal: promote progress in content-based retrieval based on end user **ad-hoc queries** that include persons, objects, locations, activities and their combinations.

- Task: Given a test collection, a query, and a master shot boundary reference, return a ranked list of at most 1,000 shots (out of 335,944) which best satisfy the need.

- New testing data: 4,593 Internet Archive videos (IACC.3), 600 total hours with video durations between 6.5 min – 9.5 min.

- Development data: ~1400 hours of previous IACC data used between 2010-2015 with concept annotations.

Query Development

- Test videos were viewed by 10 human assessors hired by NIST
- 4 facet description of different scenes were used (if applicable):
  - Who : concrete objects and being (kind of persons, animals, things)
  - What : are the objects and/or beings doing ? (generic actions, conditions/state)
  - Where : locale, site, place, geographic, architectural
  - When : time of day, season
- In total assessors watched ~35% of the IACC.3 videos
- 90 Candidate queries chosen from human written descriptions to be used between 2016-2018.

TV2016 Queries by complexity

- Person + Action + Object + Location

Find shots of a person playing guitar outdoors

Find shots of a man indoors looking at camera where a bookcase is behind him

Find shots of a person playing drums indoors

Find shots of a diver wearing diving suit and swimming under water

Find shots of a person holding a poster on the street at daytime

- Person + Action + Location

Find shots of the 43rd president George W. Bush sitting down talking with people indoors

Find shots of a choir or orchestra and conductor performing on stage

Find shots of one or more people walking or bicycling on a bridge during daytime

Find shots of a crowd demonstrating in a city street at night

TV2016 Queries by complexity

- Person + Action/state + Object

Find shots of a person sitting down with a laptop visible
Find shots of a man with beard talking or singing into a microphone
Find shots of one or more people opening a door and exiting through it
Find shots of a man with beard and wearing white robe speaking and gesturing to camera
Find shots of a person holding a knife
Find shots of a woman wearing glasses
Find shots of a person drinking from a cup, mug, bottle, or other container
Find shots of a person wearing a helmet
Find shots of a person lighting a candle

- Person + Action

Find shots of people shopping
Find shots of military personnel interacting with protesters
Find shots of soldiers performing training or other military maneuvers
Find shots of a person jumping
Find shots of a man shake hands with a woman

TV2016 Queries by complexity

- Person + Location

Find shots of one or more people at train station platform

Find shots of two or more men at a beach scene

- Person + Object

Find shots of a policeman where a police car is visible

- Object + Location

Find shots of any type of fountains outdoors

- Object

Find shots of a sewing machine

Find shots of destroyed buildings

Find shots of palm trees

Training and run types

Four training data types:

- ✓ A – used only IACC training data (4 runs)
- ✓ D – used any other training data (42 runs)
- ✓ E – used only training data collected automatically using only the query text (6 runs)
- ✓ F – used only training data collected automatically using a query built manually from the given query text (0 runs)

Two run submission types:

- ✓ Manually-assisted (M) – Query built manually
- ✓ Fully automatic (F) – System uses official query directly

Evaluation

Each query assumed to be binary: absent or present for each master reference shot.

NIST sampled ranked pools and judged top results from all submissions.

Metrics: *inferred average precision per query.*

Compared runs in terms of **mean** *inferred average precision* across the 30 queries.

mean extended Inferred average precision (xinfAP)

2 pools were created for each query and sampled as:

- ✓ Top pool (ranks 1-200) sampled at 100%
- ✓ Bottom pool (ranks 201 - 1000) sampled at 11.1%
- ✓ % of sampled and judged clips from rank 201-1000 across all runs (min= 10.5%, max = 76%, mean = 35%)

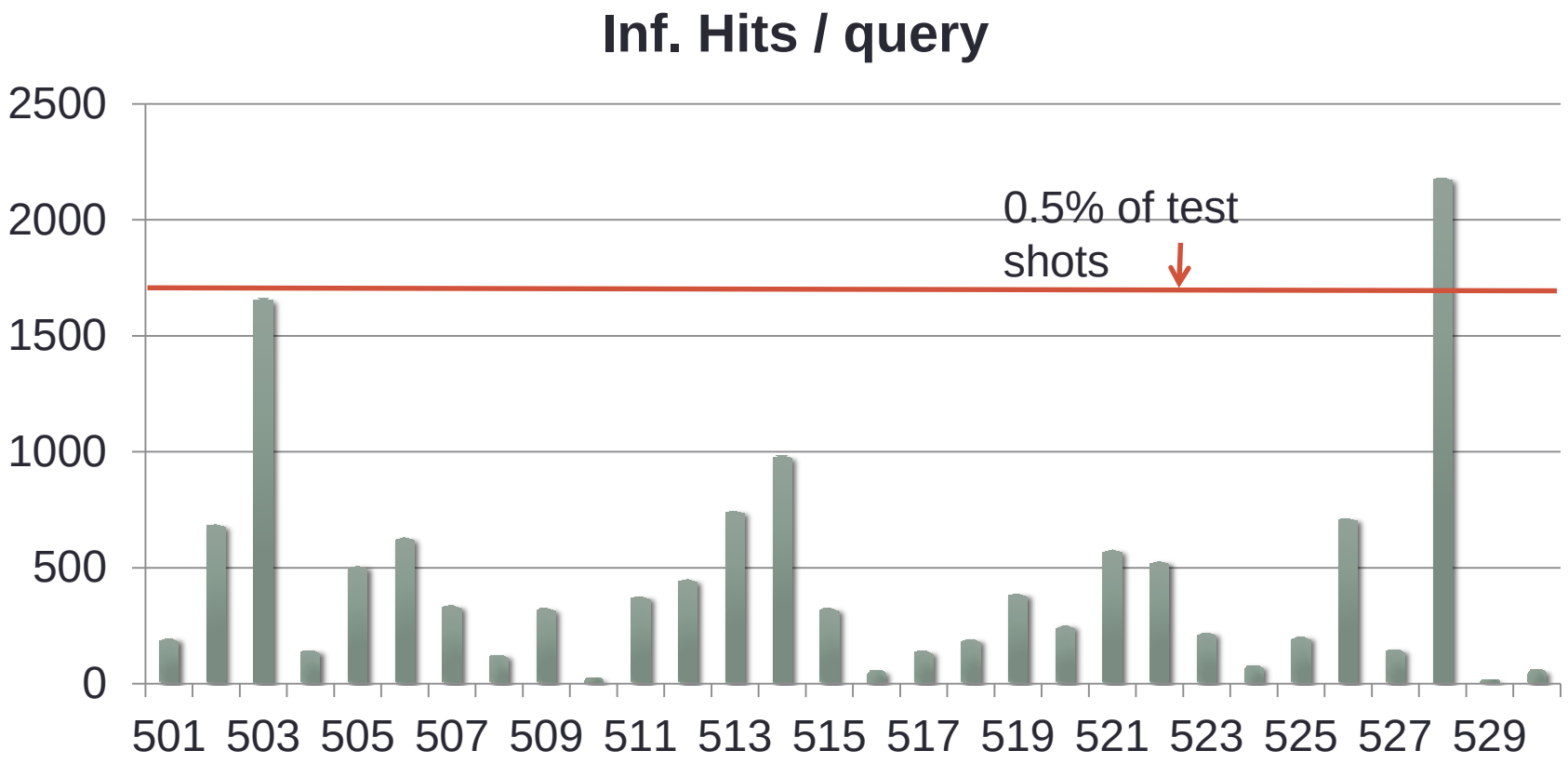
30 queries
187,918 total judgments
7,448 total hits
4642 hits at ranks (1-100)
2080 hits at ranks (101-200)
726 hits at ranks (201-2000)

Judgment process: one assessor per query, watched complete shot while listening to the audio. infAP was calculated using the judged and unjudged pool by sample_eval

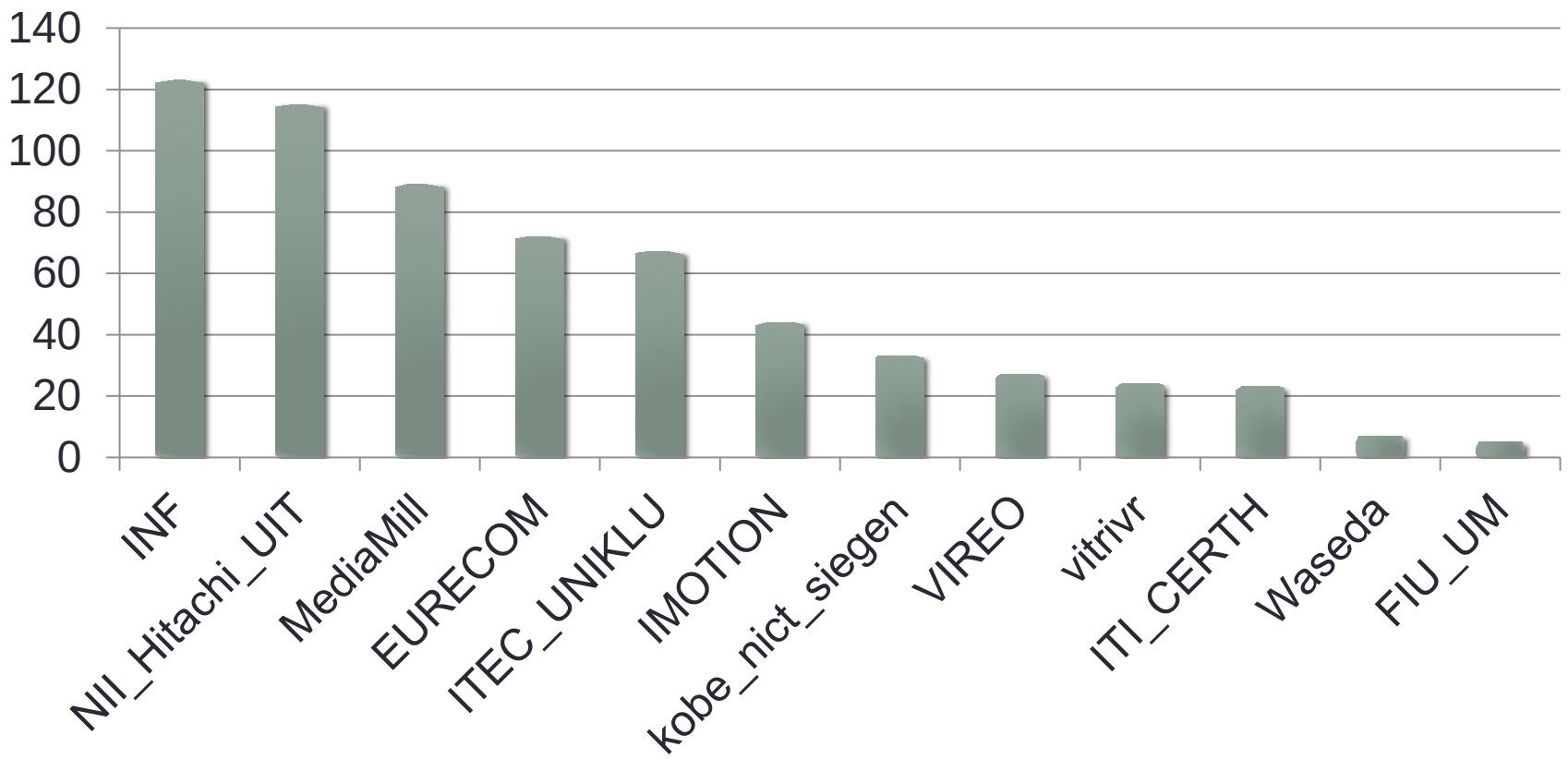
Finishers : 13 out of 29

		M	F
INF	CMU; Beijing U. of Posts and Telecommunication; U. Autonoma de Madrid; Shandong U.; Xian JiaoTong U. Singapore	-	4
kobe_nict_siegen	Kobe U.; Japan National Institute of Information and Communications Technology, Japan U. of Siegen, Germany	3	-
UEC	Dept. of Informatics, The U. of Electro-Communications, Tokyo	2	-
ITI_CERTH	Inf. Tech. Inst., Centre for Research and Technology Hellas	4	4
ITEC_UNIKLU	Klagenfurt U.	-	3
NII_Hitachi_UIT	Natl. Inst. Of Info.; Hitachi Ltd; U. of Inf. Tech.(HCM-UIT)	-	4
IMOTION	U. of Basel, Switzerland; U. of Mons, Belgium; Koc U., Turkey	2	2
MediaMill	U. of Amsterdam Qualcomm	-	4
Vitrivr	U. of Basel	2	2
Waseda	Waseda U.	4	-
VIREO	City U. of Hong Kong	3	3
EURECOM	EURECOM	-	4
FIU_UM	Florida International U., U. of Miami	2	-

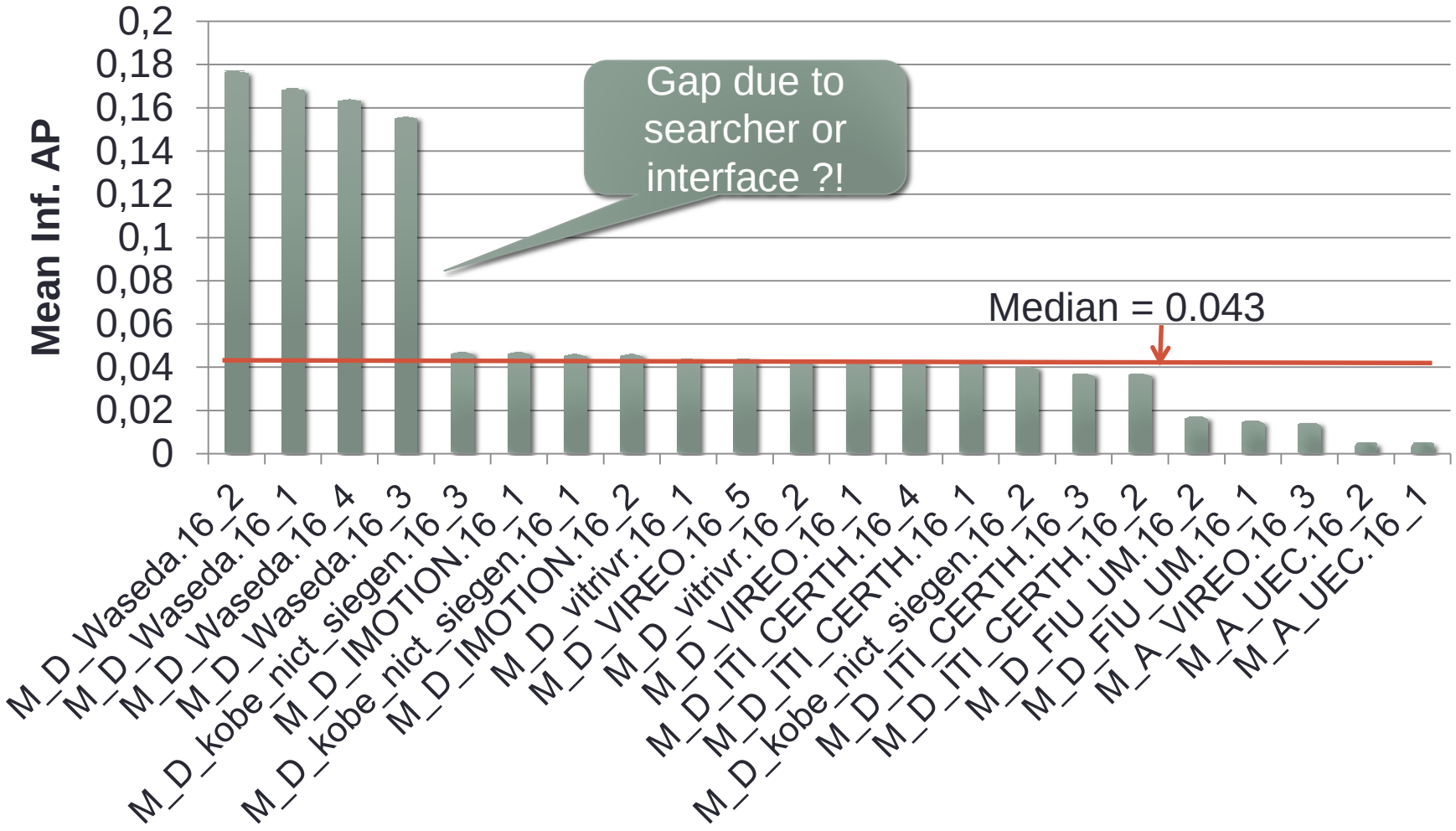
Inferred frequency of hits varies by query



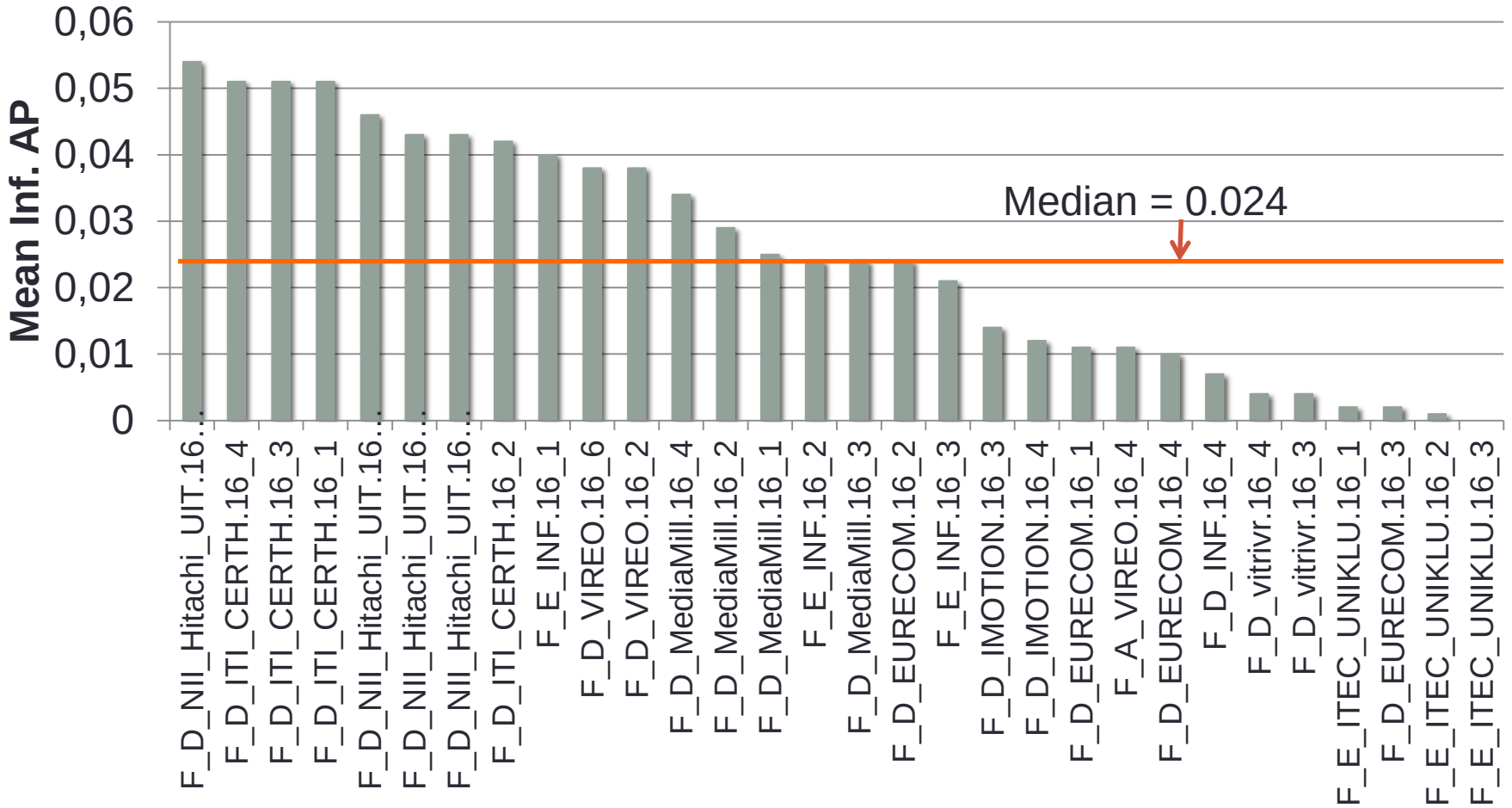
Total true shots contributed uniquely by team



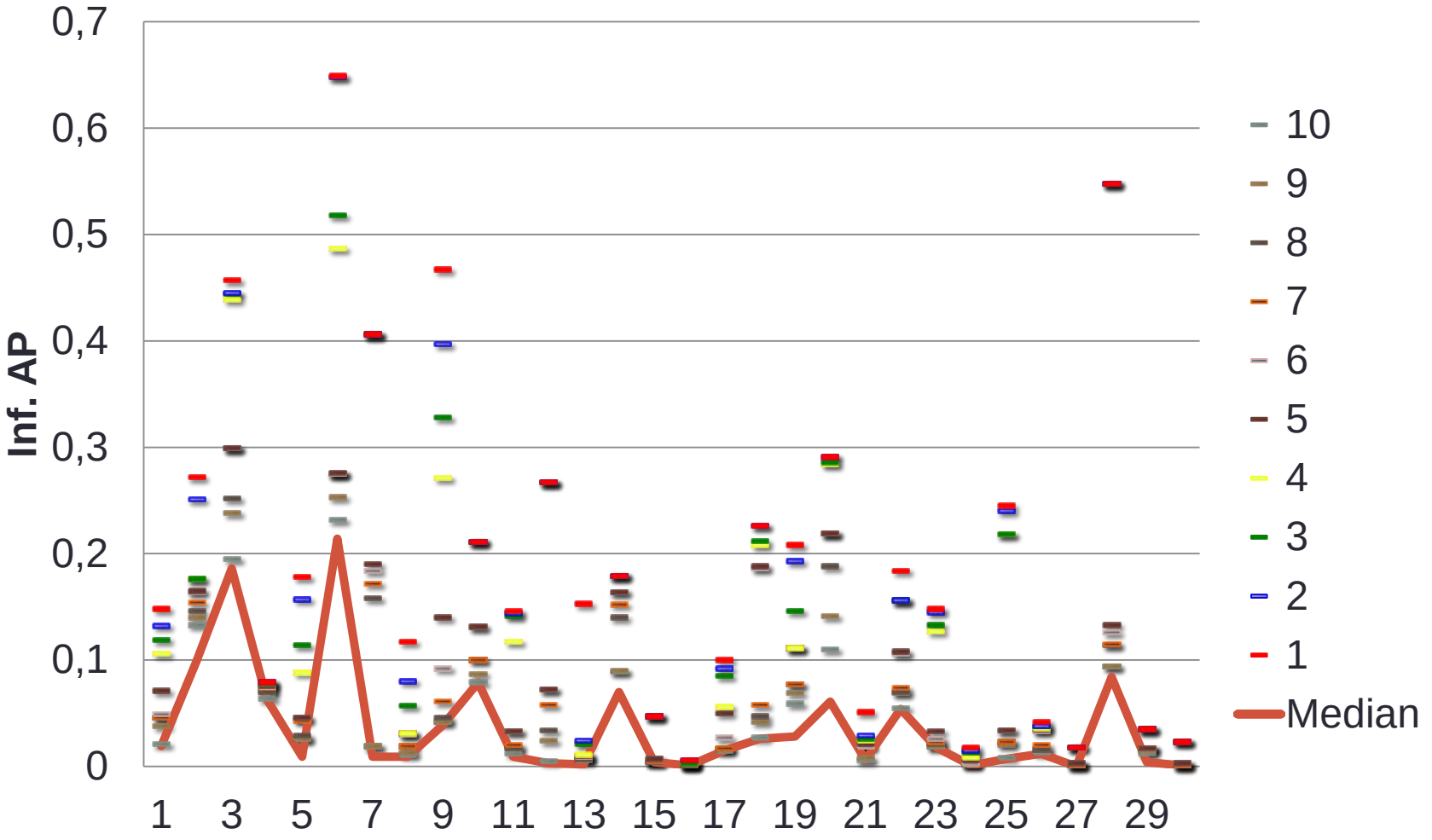
2016 run submissions scores (22 Manually-assisted runs)



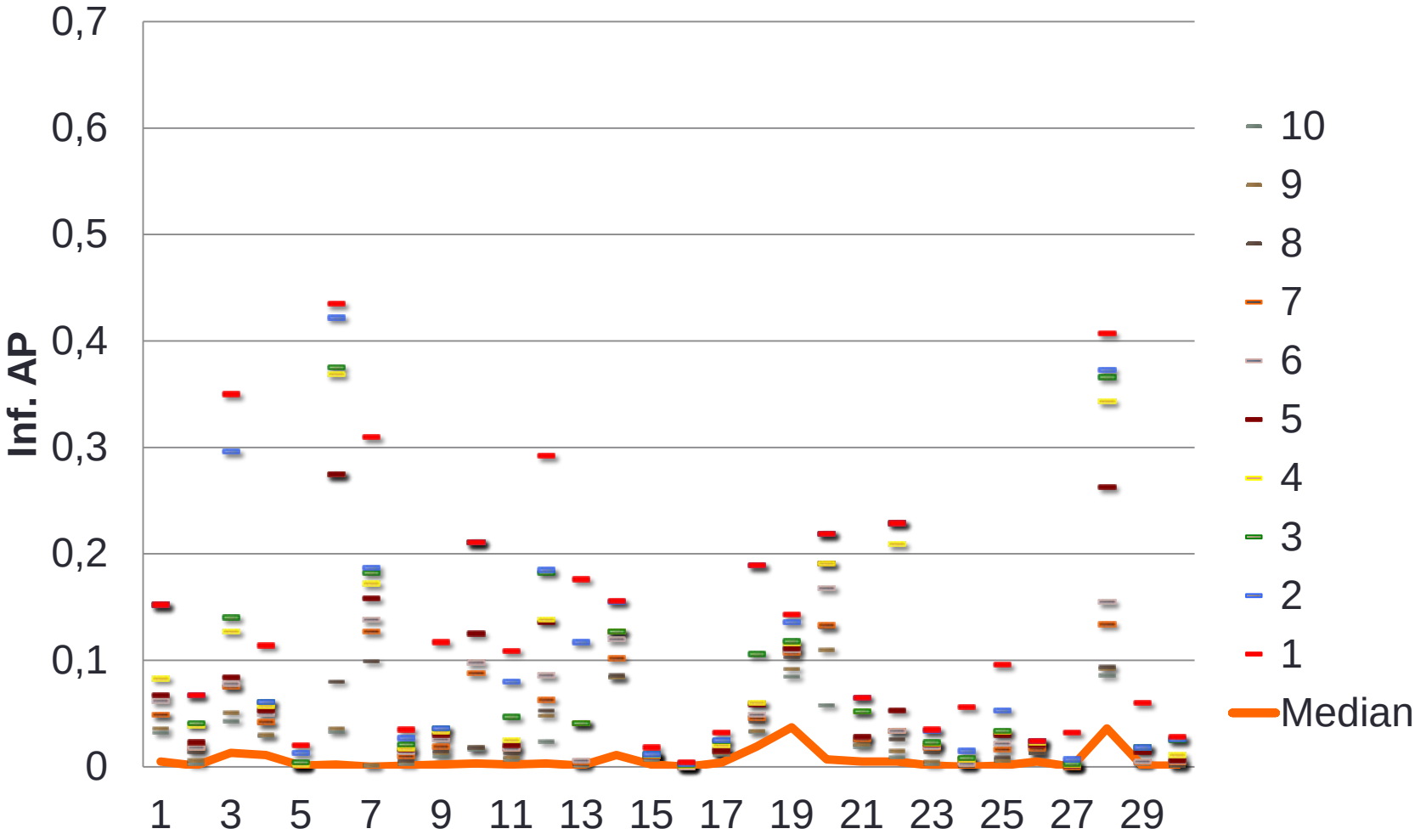
2016 run submissions scores (30 Fully automatic runs)



Top 10 infAP scores by query (Manually-assisted)



Top 10 infAP scores by query (Fully automatic)



Statistical significant differences among top 10 "M" runs (using randomization test, p < 0.05)

D_Waseda.16_2
- D_Waseda.16_3
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5
- D_Waseda.16_4
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5

D_Waseda.16_1
- D_Waseda.16_3
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5

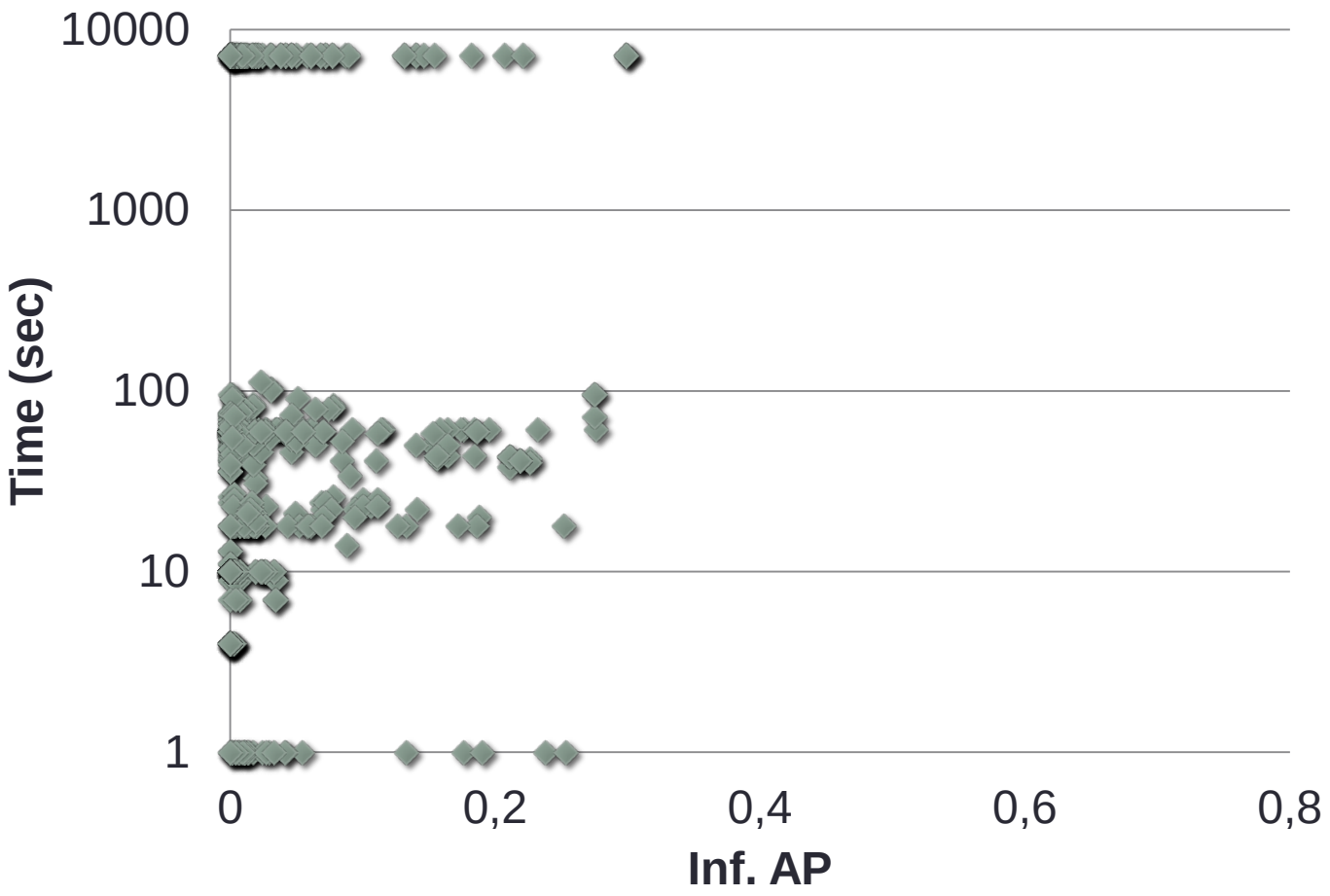
Run	Inf. AP score
D_Waseda.16_2	0.177 *
D_Waseda.16_1	0.169 *
D_Waseda.16_4	0.164 #
D_Waseda.16_3	0.156 #
D_kobe_nict_siegen.16_3	0.047 ^
D_IMOTION.16_1	0.047 ^
D_kobe_nict_siegen.16_1	0.046 ^
D_IMOTION.16_2	0.046 ^
D_vitrivr.16_1	0.044 ^
D_VIREO.16_5	0.044 ^

Statistical significant differences among top 10 “F” runs (using randomization test, p < 0.05)

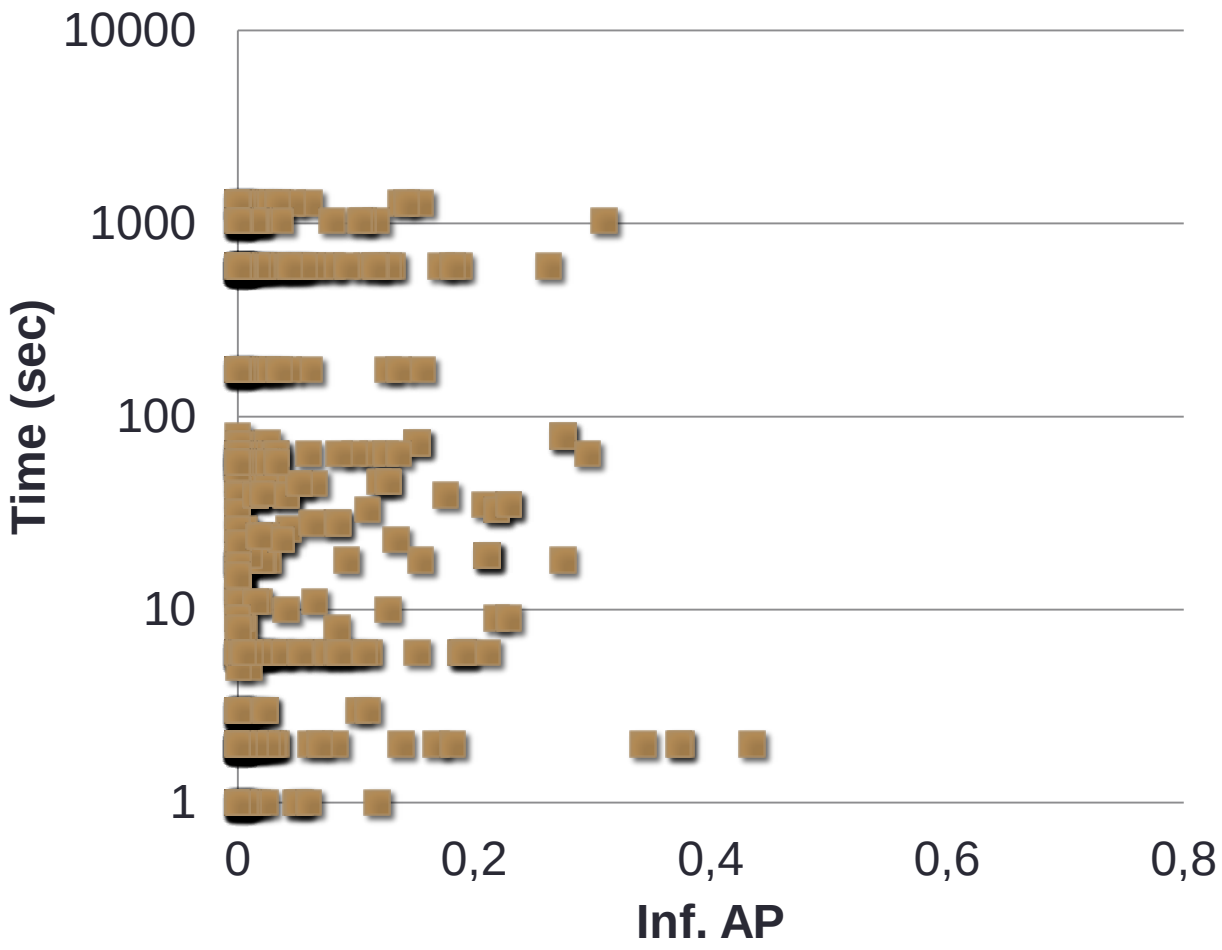
Run	Inf. AP score
D_NII_Hitachi_UIT.16_4	0.054
D_ITI_CERTH.16_4	0.051
D_ITI_CERTH.16_3	0.051
D_ITI_CERTH.16_1	0.051
D_NII_Hitachi_UIT.16_3	0.046
D_NII_Hitachi_UIT.16_2	0.043
D_NII_Hitachi_UIT.16_1	0.043
D_ITI_CERTH.16_2	0.042
E_INF.16_1	0.040
D_VIREO.16_6	0.038

No statistical significant differences among the top 10 runs

Processing time vs Inf. AP (“M” runs)



Processing time vs Inf. AP (“F” runs)



2016 Observations

- Most teams relied on intensive visual concept indexing, leveraging on past SIN task and similar like ImageNet, Scenes …
- Combined with manual or automatic query transformation
- Clever combination of concept scores (e.g. Waseda)

- Ad-hoc search is more difficult than simple concept-based tagging.
- Big gap between SIN best performance and AVS: maybe performance should be better compared with the “concept pair” task within SIN

- Manually-assisted runs performed better than fully-automatic.
- Most systems are not real-time (slower systems were not necessarily effective).
- E and F runs are still rare compared to A and D

Continued at MMM2017

- 10 Ad-Hoc Video Search (AVS) tasks, 5 of which are a random subset of the 30 AVS tasks of TRECVID 2016 and 5 will be chosen directly by human judges as a surprise. Each AVS task has several/many target shots that should be found.
- 10 Known-Item Search (KIS) tasks, which are selected completely random on site. Each KIS task has only one single 20-seconds long target segment

PART II
Some participants' implementations

Papers on the NIST server:

http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.16.org.html

General approach

- Gather / develop “concept banks”
  - Lists of concepts with associated detectors
- Match query elements to available concepts
  - Manually or automatically select concepts in the lists with the query elements
  - Use concept names and definitions
- Score and sort shots according to the selected concepts’ scores and the query
  - Combine the scores of the selected concepts

Concept banks

- Typically DCNN or SVM classifiers trained on annotated image or video collections
- Many teams uses “off the shelf” state of the art and publicly available implementations (e.g. caffe “model zoo”)
- Precomputed detection scores for a number of models × concept lists
- These may be used for a number of tasks beyond AVS, eg. TRECVid MED or NTCIR lifelog
- In case of full videos or shots: max pooling on multiple frames

Concept banks

- Many concept lists contain exclusive classes, e.g. ImageNet LSVRC
  - OK for the target collection of typical samples (that may contain either a cat or a dog but not both)
  - NOT OK for samples “from the wild” (that may contain both a cat and a dog for instance)
  - Generally remove the soft-max output layer
    - Better if concepts are not exclusive
    - Better for MAP metrics
- Many pre-trained models may be available for a same concept list, e.g. ImageNet LSVRC
  - Normalize and fuse (average) predicted scores

Most popular concept banks

- ImageNet LSVRC: 1000 exclusive classes, many pre-trained models
- Places-205/365: 205/365 exclusive classes of places, pre-trained models
- Hybrid ILSVRC / places models
- ImageNet shuffle: lists of 1000, 4437, 8201, 12988 and 4000 concepts, pre-trained models
- TRECVid 346: SIN task concepts, many pre-trained models or fine-tuned models
- FCVID 239: activities, events
- Sports 487: activities, events

Matching query elements to concepts

- Difficult because of:
  - Polysemy / synonymy
  - Not so good concept names and definitions
  - What is necessary may not be in the available concepts
- Use of NLP techniques like semantic similarity
- User interfaces for making proposal and letting the user select the appropriate concepts
- Use generic, specific or related concepts in the absence of exact matches (always better than chance)

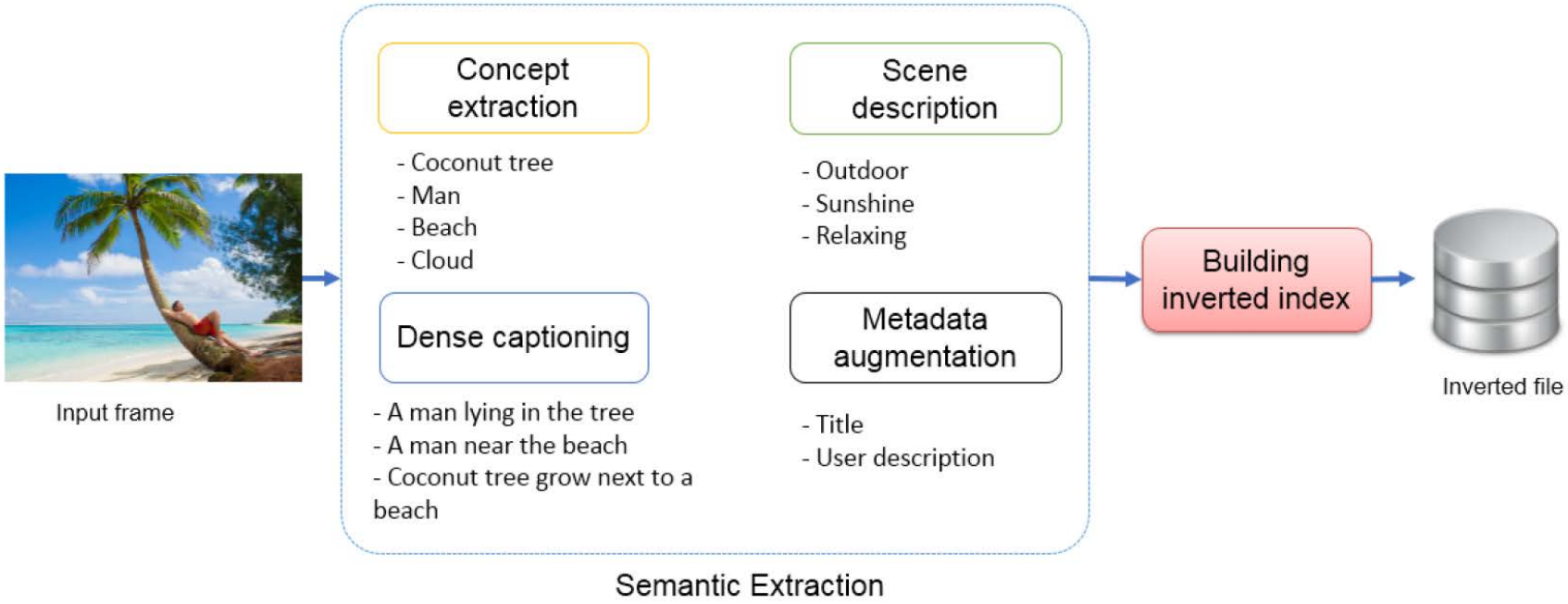
Scoring shots from concept scores

- Query decomposed into “required elements”
- For each element:
  - One or more concepts have been found
  - One or more scores available for each concept
  - Normalize and average everything for robustness for producing a score for the element
- Simple case (most common)
  - Combine the scores of the required elements at the query level using an associative operator (sum, product, min …)
- More elaborate
  - Consider Boolean expressions of the elements

Other approaches (NII-Hitachi-UIT)

- Build an inverted index
- Associate words to the video shots using
  - previous methods and thresholding
  - Image captioning techniques
- Use then classical text IR techniques (Lucene)

Other approaches (NII-Hitachi-UIT)



Use of Boolean expressions

- Waseda university AVS system
- Manually assisted submissions
- With kind permission of Kazuya UEKI

Kazuya UEKI, Kotaro KIKUCHI, Susumu SAITO and Tetsunori KOBAYASHI. ***Waseda at TRECVID 2016; Ad-hoc Video Search(AVS).*** TRECVID 2016, NIST, USA.

http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/waseda.pdf

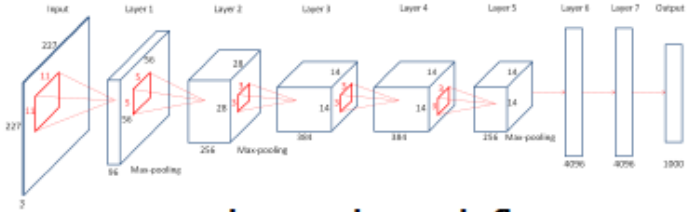
2. System description

Our method consists of three steps:

[Step. 1]

Manually select several search keywords based on the given query phrase.

[Step. 2]



Calculate a score for each concept using visual features.

[Step. 3]

Combine the semantic concepts to get the final scores.

2. System description

[Step. 1]

Manually select several search keywords based on the given query phrase.

We explicitly distinguished *and* from *or*.

Example 1

"any type of fountains outdoors"

⇒ "fountain" and "outdoor"

Example 2

"one or more people walking or bicycling on a bridge during daytime"

⇒ "people" and ("walking" or "bicycling") and "bridge" and "daytime"

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

We extracted visual features from pre-trained convolutional neural networks (CNNs)

Pre-trained models used in our runs

Model name	Database	Number of concepts	Concept type(s)
TRECVID346	TRECVID (ImageNet)	346	Object, Scene, Action
PLACES205	Places	205	Scene
PLACES365	Places	365	Scene
HYBRID1183	Places, ImageNet	1,183	Object, Scene
IMAGENET1000	ImageNet	1,000	Object
IMAGENET4437	ImageNet	4,437	Object
IMAGENET8201	ImageNet	8,201	Object
IMAGENET12988	ImageNet	12,988	Object
IMAGENET4000	ImageNet	4,000	Object

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

Score normalization

The score for each semantic concept was normalized over all the test shots such that the maximum and the minimum scores were 1.0 (most probable) and 0.0 (least probable).

Concept selection

No concept name matching a given search keyword.

⇨ Semantically similar concept was chosen by word2vec.

Search keyword did not have a semantically similar concept.

⇨ This keyword was not used.

2. System description

[Step. 3]

Combine the semantic concepts to get the final scores.

Score fusion

Calculate the final scores by score-level fusion

or operator

"walking" or "bicycling"

0.40 0.10 → maximum score → 0.40

and operator

"fountain" and "outdoor"

0.90 0.80

summing score → 0.90 + 0.80 = 1.70

(*) depend on runs

multiplying score → 0.90 x 0.80 = 0.72

3. Submission

Waseda1 run

Total score was simply calculated by multiplying the scores of the selected concepts.

$$\prod_{i=1}^{N} S_i$$

N: # selected concepts

S_i: normalized score

"fountain" and "outdoor"

shot A:	0.70	x	0.10	=	0.07
shot B:	0.40	x	0.30	=	0.12
	⋮		⋮		⋮

Shots having all the selected concepts will tend to appear in the higher ranks.

3. Submission

Waseda2 run

Almost the same as Waseda1 run except for the incorporation of a fusion weight.

$$\prod_{i=1}^{N} S_i^{w_i}$$

fusion weight (= IDF values) calculated from the Microsoft COCO database.

A rare keyword is of higher importance than an ordinary keyword.

	"man"	and	"bookcase"
shot A:	$(0.90)^{1.97}$	x	$(0.70)^{8.23}$
	= 0.81	x	0.05 = 0.04
shot B:	$(0.70)^{1.97}$	x	$(0.90)^{8.23}$
	= 0.50	x	0.42 = 0.21

3. Submission

Waseda3 run

Total score was calculated by summing the scores of the selected concepts.

$$\sum_{i=1}^{N} S_i$$

	“fountain” and “outdoor”				
shot A:	0.70	+	0.10	=	0.80
shot B:	0.40	+	0.30	=	0.70
	⋮		⋮		⋮

Somewhat looser conditions than multiplying (Waseda1, Waseda2 runs)

3. Submission

Waseda4 run

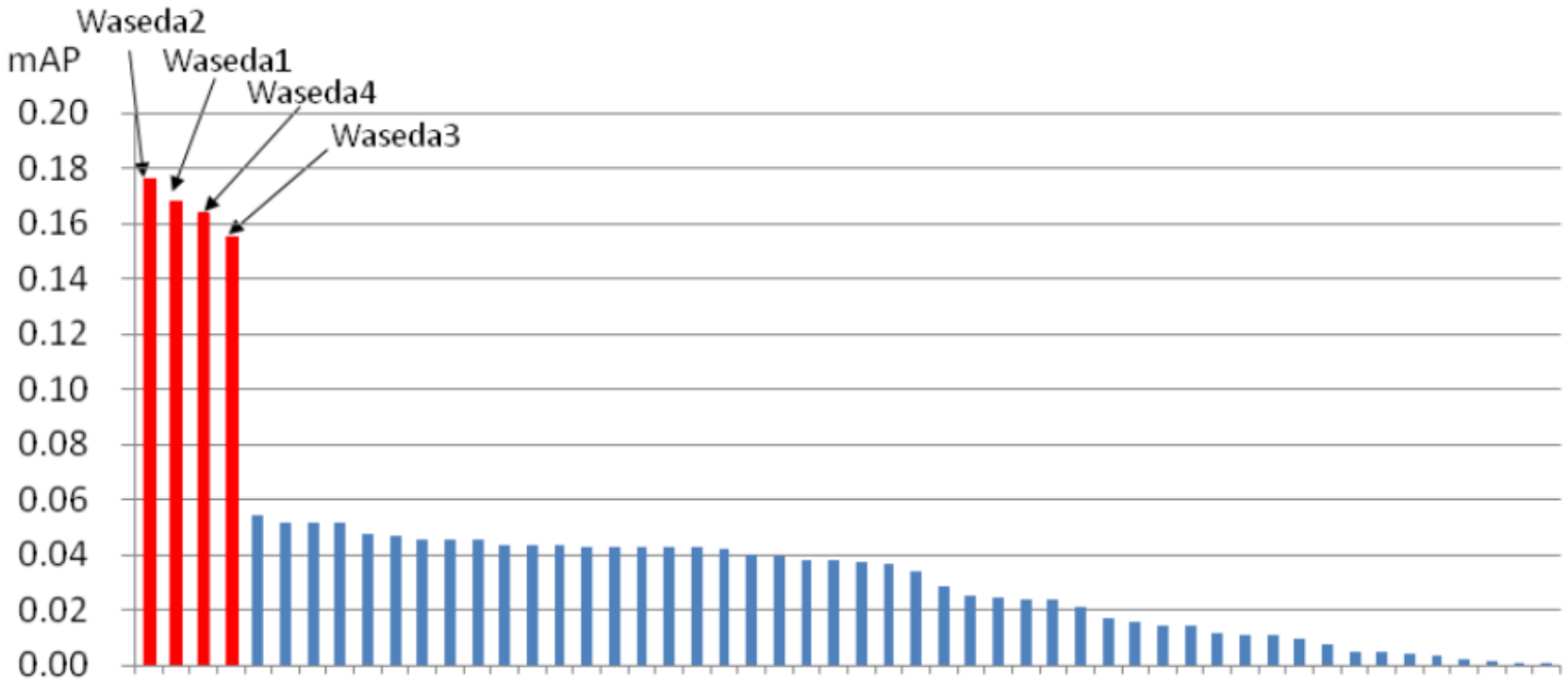
Similar to Waseda3 except that fusion weight is used.

$$\sum_{i=1}^{N} w_i \cdot s_i$$

"man" and "bookcase"

shot A:	(1.97 x 0.90)	+	(8.23 x 0.70)	=	7.53
shot B:	(1.97 x 0.70)	+	(8.23 x 0.90)	=	8.79

4. Results



Comparison of Waseda runs with the runs of other teams on IACC_3

Our 2016 submissions ranked between 1 and 4 in a total of 52 runs. Our best run was a mean average precision of 17.7%.

4. Results

Comparison of Waseda runs

Name	Fusion method	Fusion weight	mAP
Waseda1	Multiplying scores		16.9
Waseda2	Multiplying scores	✓	17.7
Waseda3	Summing scores		15.6
Waseda4	Summing scores	✓	16.4

- The stricter condition in which all the concepts in a query phrase must be included has the better performance.
- The rarely seen concepts are much more important for the video retrieval task.

Conclusion

- AVS becomes a feasible task for a number of non-trivial cases
- Concept (including objects, scenes, actions …) are the main basis for most systems
- Still more concepts are needed
- Boolean expressions boost performance (though a contrast experiment is missing)
- Otherwise fully automatic systems are on par with manually assisted ones
- Most future work is likely to be in automatic topic / query processing using NLP and QA type techniques

Thanks

Slides available from:

http://mrim.imag.fr/georges.quenot/icmr2017/AVS.pdf